Information Processing Letters 111 (2011) 895–898

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

# Lower bounds for context-free grammars

Yuval Filmus [1]

*University of Toronto, Canada*

ARTICLE INFO

*Article history:*
Received 23 February 2011
Received in revised form 14 June 2011
Accepted 15 June 2011
Available online 1 July 2011
Communicated by J. Torán

*Keywords:*
Formal languages
Context-free grammars
Lower bounds

ABSTRACT

Ellul, Krawetz, Shallit and Wang prove an exponential lower bound on the size of any context-free grammar generating the language of all permutations over some alphabet. We generalize their method and obtain exponential lower bounds for many other languages, among them the set of all squares of given length, and the set of all words containing each symbol at most twice.



## 1. Introduction

How efficiently can we represent a given set of strings using a context-free grammar? We show that for many simple languages, any context-free grammar must have size $\Omega(c^n)$ for some constant $c$, where $n$ is some natural parameter (the size of the alphabet or the size of the words in question). Examples are the set of all permutations over some alphabet of size $n$, the set of all squares $w^2$ of size $2n$ over some fixed alphabet, and the set of all words over an alphabet of size $n$ containing each symbol exactly (or at most) $k$ times.

Our method generalizes the method used by Ellul, Krawetz, Shallit and Wang [1] to prove an exponential lower bound on the size of context-free grammars generating the set of all permutations over a finite alphabet.

A similar question has been considered by Charikar et al. [2] and Arpe and Reischuk [3], who show that it is hard to approximate the size of the smallest grammar generating a given word.

Asveld presents several grammars for generating the set of all permutations over some alphabet [4,5], as well as the set of all cyclic shifts of some given word [6,7].

## 2. Definitions

The cardinality of a set $S$ will be denoted by $\#S$.

The set of natural numbers *including zero* will be denoted by $\mathbb{N}$.

The set of all permutations over a set $A$ will be denoted by $\mathcal{S}(A)$.

We will use $\mathbb{A}_t$ to denote some fixed alphabet of cardinality $t$. A *word* over $\mathbb{A}_t$ is a (possibly empty) sequence of symbols from $\mathbb{A}_t$. The length of a word $w$, denoted by $|w|$, is the number of symbols in $w$.

A word $x$ is a (consecutive) *subword* of $w$, denoted by $x \trianglelefteq w$, if $w = lxr$ for some (possibly empty) words $l, r$.

For a language $L$, denote by $\mathrm{sw}(L)$ the set of all subwords of words in $L$, that is

$$\mathrm{sw}(L) = \{x\colon\ x \trianglelefteq w \text{ for some } w \in L\}.$$

We shall use $G = (N, T, P, S)$ for a context-free grammar, where $N$ is the set of non-terminals, $T$ is the set of terminals, $P$ is the set of productions, and $S$ is the start symbol.

Following Kelemenová [8], we define the size of a production $A \to \alpha$ as $|\alpha| + 2$. The size of a context-free grammar $G = (N, T, P, S)$, denoted by $\mathrm{size}(G)$, is the sum of the sizes of all productions in $P$.

A context-free grammar $G$ is said to be in *Chomsky normal form* if every production of $G$ is of one of the forms

*E-mail address:* yuvalf@cs.toronto.edu.

[1] Supported by NSERC.

0020-0190/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ipl.2011.06.006

$A \to BC, \quad A \to a, \quad S \to \epsilon,$

where $A, B, C$ are non-terminals, $a$ is a terminal, and $S$ is the start symbol.

The following theorem is well known.

**Theorem 1.** *For every context-free grammar $G$ there exists a context-free grammar in Chomsky normal form of size $O(\text{size}(G)^2)$ generating the same language.*

## 3. Method

Grammars in Chomsky normal form satisfy the following well-known *subword lemma*, which is the key to our method.

**Lemma 2.** *Suppose the word $w$ is generated by a context-free grammar $G$ in Chomsky normal form, and furthermore $|w| \geqslant 2$. For each positive $\ell \leqslant |w|$ there is a subword $x$ of $w$ of length $\ell/2 \leqslant |x| < \ell$ generated by a non-terminal of $G$.*

**Proof.** Consider the derivation tree of $w$. For every node $v$ in the tree, denote by $\|v\|$ the size of the subword of $w$ generated by $v$.

Define a sequence of nodes $v_1, v_2, \ldots, v_k$ in the derivation tree inductively as follows. The first vertex $v_1$ is the root of the derivation tree. If $v_i$ is a node in the tree that has one child (which must be a terminal), the sequence terminates. If $v_i$ is a node that has two children, arrange its children $x, y$ so that $\|x\| \geqslant \|y\|$, and let $v_{i+1} = x$.

Consider the first node $v_i$ such that $\|v_i\| < \ell$; such a node exists since $\|v_k\| = 1$. Since $\|v_1\| \geqslant \ell$, necessarily $i > 1$, and so $\|v_{i-1}\| \geqslant \ell$. Our rule for choosing $v_i$ implies that $\|v_i\| \geqslant \|v_{i-1}\|/2 \geqslant \ell/2$. $\square$

Our method makes use of a complexity measure $M$ defined as follows.

**Definition 3.** Let $L$ be a context-free language, $\ell \geqslant 2$ an integer, and $W$ a subset of $L$, all of whose words are of length at least $\ell$.

Define a language $X$ as follows:

$$X = \{x\colon\ \ell/2 \leqslant |x| < \ell\} \cap \text{sw}(W).$$

Define a reflexive, symmetric relation $\sim$ on $X$ by letting $x \sim y$ if there exist words $\alpha, \beta, \gamma, \delta$ such that

$$\alpha x\beta, \gamma y\delta \in W, \quad \alpha y\beta, \gamma x\delta \in L.$$

A subset $C \subset X$ is a *clique* if $x \sim y$ for all $x, y \in C$.

For any subset $C \subset X$, define its *complexity* $M(C)$ by

$$M(C) = \#\{w \in W\colon\ x \trianglelefteq w \text{ for some } x \in C\}.$$

In words, $M(C)$ is the number of words in $W$ that have some subword in $C$.

Finally, define $M(L, \ell, W)$ as the maximum of $M(C)$ over all cliques $C$.

In all the applications below, $\sim$ will be an equivalence relation, and so instead of cliques we can consider equivalence classes in Definition 3.

**Lemma 4.** *Let $L$ be a context-free language, $\ell \geqslant 2$ an integer, and $W$ a subset of $L$, all of whose words are of length at least $\ell$.*

*If the relation $\sim$ defined in Definition 3 is an equivalence relation, then $M(L, \ell, W)$ is equal to the maximum of $M(C)$ over all equivalence classes $C$.*

**Proof.** If $\sim$ is an equivalence relation then any clique is a subset of some equivalence class. The lemma follows from the monotonicity of $M(C)$. $\square$

In most applications we will have $W = L$. In that case we can simplify the definition of the relation $\sim$.

**Lemma 5.** *Let $\ell \geqslant 2$ be an integer, and $L$ be a context-free language consisting of words of length at least $\ell$.*

*Let $X$ be the set in Definition 3, where we set $W = L$. Define a reflexive, symmetric relation $\approx$ on $X$ by letting $x \approx y$ if there exist words $\alpha, \beta$ such that*

$$\alpha x\beta, \alpha y\beta \in L.$$

*The relation $\approx$ coincides with the relation $\sim$ defined in Definition 3.*

**Proof.** Let $x, y \in X$. If $x \sim y$ then there exist words $\alpha, \beta$ such that $\alpha x\beta \in W = L$ and $\alpha y\beta \in L$. Thus $x \approx y$.

Conversely, if $x \approx y$ then there exist words $\alpha, \beta$ such that $\alpha x\beta, \alpha y\beta \in L$. Letting $\gamma = \alpha$ and $\delta = \beta$, we see that also $x \sim y$. $\square$

Our method is summarized by the following proposition.

**Proposition 6.** *Let $L$ be a context-free language, $\ell \geqslant 2$ an integer, and $W$ a subset of $L$, all of whose words are of length at least $\ell$.*

*Let $M = M(L, \ell, W)$ be the parameter defined in Definition 3. Every context-free grammar for $L$ has size*

$$\Omega\left(\sqrt{\frac{\#W}{M}}\right).$$

**Proof.** We show that every context-free grammar $G$ in Chomsky normal form which generates $L$ contains at least $\#W/M$ non-terminals. The proposition follows from Theorem 1.

Let $G$ be a context-free grammar $G$ in Chomsky normal form which generates $L$. Using Lemma 2, we can associate with each $w \in W$ a subword $x(w) \in X$ generated by some non-terminal $N(w)$. For a non-terminal $A$, let

$$N^{-1}(A) = \{w \in W\colon\ N(w) = A\}.$$

Suppose that $w_1, w_2 \in N^{-1}(A)$. Write $w_1 = \alpha x(w_1)\beta$, $w_2 = \gamma x(w_2)\delta$. Note that $A$ generates both $x(w_1)$ and $x(w_2)$, and so $\alpha x(w_2)\beta, \gamma x(w_1)\delta \in L$. In other words, $w_1 \sim w_2$. We conclude that $N^{-1}(A)$ is a clique. By the definition of $M$,

$$\#N^{-1}(A) \leqslant M\big(N^{-1}(A)\big) \leqslant M.$$

Since the sets $N^{-1}(A)$ form a partition of $W$ into parts of cardinality at most $M$, we deduce that $G$ must contain at least $\#W/M$ non-terminals. $\square$

## 4. Applications

We now present several applications of Proposition 6. The first application concerns the language of all squares of words of a given length.

**Theorem 7.** *Let $t \geqslant 2$ be an integer and $L = \{w^2\colon\ w \in \mathbb{A}_t^n\}$.*

*Every context-free grammar for $L$ has size*

$$\Omega\left(\frac{t^{n/4}}{\sqrt{2n}}\right).$$

**Proof.** We use the following definition: the *root* of a word $w^2 \in L$ is defined to be $w$.

Let $W = L$ and $\ell = n$ in Definition 3. Suppose $x, y \in X$ and $\alpha x \beta, \alpha y \beta \in L$. Since $|x|, |y| \leqslant n$, the root of a word of the form $\alpha z \beta \in L$ is recoverable from $\alpha$ and $\beta$, so that $x = y$. Thus each clique consists of a single word.

We now estimate the number of words in $L$ containing a given $x \in X$ as a subword. There are fewer than $2n$ possible starting locations for $x$. For each starting location, $x$ determines $|x| \geqslant n/2$ symbols of the root, and so there are at most $t^{n/2}$ possible roots. In total, at most $M = 2nt^{n/2}$ words in $L$ contain any given $x \in X$. Since $\#L = t^n$, we have $\#L/M = t^{n/2}/2n$, and the theorem follows from Proposition 6. $\square$

The next application generalizes Theorem 7 to the language of all $k$th powers for $k \geqslant 3$. Moreover, we allow an arbitrary permutation to be applied on each of the $k$ copies.

**Theorem 8.** *Let $t \geqslant 2$, $n \geqslant 2$ and $k \geqslant 3$ be integers, and $\pi_1, \ldots, \pi_k \in \mathcal{S}(\mathbb{A}_t^n)$ be permutations. Let $L = \{\pi_1(w) \cdots \pi_k(w)\colon\ w \in \mathbb{A}_t^n\}$.*

*Every context-free grammar for $L$ has size*

$$\Omega\left(\frac{t^{n/8}}{\sqrt{kn}}\right).$$

**Proof.** We use the following definition: the *root* of a word $\pi_1(w) \cdots \pi_k(w) \in L$ is defined to be $w$.

Let $W = L$ and $\ell = n$ in Definition 3. Suppose $x, y \in X$ and $\alpha x \beta, \alpha y \beta \in L$. Since $|x| \leqslant n$, either $\alpha$ contains $\pi_1(w)$ or $\beta$ contains $\pi_k(w)$, where $w$ is the root of $\alpha x \beta$ (here we use $k \geqslant 3$). Since the $\pi_i$ are permutations, we get that $\alpha x \beta$ and $\alpha y \beta$ have the same root, and so $x = y$. Thus each clique consists of a single word.

We now estimate the number of words in $L$ containing a given $x \in X$ as a subword. There are fewer than $kn$ starting locations for $x$. For each starting location, $x$ intersects the location of some $\pi_i$ in at least $|x|/2 \geqslant n/4$ points. Thus for each starting location, there are at most $t^{3n/4}$ possible roots. In total, at most $M = knt^{3n/4}$ words in $L$ contain any given $x \in X$. Since $\#L = t^n$, we have $\#L/M = t^{n/4}/kn$, and the theorem follows from Proposition 6. $\square$

Note that the condition $k \geqslant 3$ in Theorem 8 is crucial: if we take $\pi_1$ as the identity and $\pi_2$ as word reversal, there is a grammar for $\{\pi_1(w)\pi_2(w)\colon\ w \in \mathbb{A}_t^n\}$ of size $O(nt)$.

The final application generalizes Theorem 30 in Ellul et al. [1].

**Theorem 9.** *Let $t \geqslant 2$, and $\Lambda \subset \mathbb{N}$ be an arbitrary subset different from $\varnothing, \{0\}, \mathbb{N}$. Let $L$ consist of all words over $\mathbb{A}_t$ in which the number of occurrences of every symbol is in $\Lambda$.*

*If $L$ is context-free then every context-free grammar for $L$ has size*

$$\Omega\left(\frac{(3^{1/2}/2^{1/3})^n}{t^{3/4}}\right).$$

**Proof.** It is easy to see that our assumptions on $\Lambda$ imply the existence of some non-zero $k \in \Lambda$ such that either $k-1 \notin \Lambda$ or $k+1 \notin \Lambda$; denote the latter element $k' \notin \Lambda$.

Let $W = \{w^k\colon\ w \in \mathcal{S}(A)\}$ and $\ell = 2t/3$ in Definition 3. Note that $\#W = t!$. We call $w$ the *root* of $w^k \in W$. Suppose $x, y \in X$ and

$$\alpha x \beta, \gamma y \delta \in W, \quad \alpha y \beta, \gamma x \delta \in L. \tag{1}$$

Since $|x|, |y| \leqslant \ell \leqslant t$, the words $x$ and $y$ contain each element $a \in \mathbb{A}_t$ at most once. Denote by $N_a(z)$ the number of occurrences of $a \in \mathbb{A}_t$ in a word $z$. The conditions (1) imply

$$\begin{aligned}
&N_a(\alpha) + N_a(x) + N_a(\beta) = k,\\
&N_a(\gamma) + N_a(y) + N_a(\delta) = k,\\
&N_a(\alpha) + N_a(y) + N_a(\beta) \neq k',\\
&N_a(\gamma) + N_a(x) + N_a(\delta) \neq k'.
\end{aligned}$$

These equations imply that $|N_a(x) - N_a(y)| \neq |k' - k| = 1$, and since $N_a(x), N_a(y) \in \{0, 1\}$, we see that $N_a(x) = N_a(y)$. Thus $x \sim y$ if and only if $y$ is a permutation of $x$. We deduce that $\sim$ is an equivalence relation, and that each equivalence class consists of all permutations over some subset $B$ of $\mathbb{A}_t$ of cardinality $t/3 \leqslant \#B < 2t/3$.

We proceed to estimate the number of words in $W$ containing a subword $x$ which is a permutation of some subset $B$. For each starting location, $x$ determines $\#B$ symbols of the root; the part of the root which is determined depends only on the starting location of $x$ modulo $t$, for which there are $t$ possibilities. Thus the number of words in $W$ containing a subword which is a permutation of $B$ is at most

$$M = t(\#B)!(t - \#B)!.$$

It is well known that the binomial coefficients $\binom{t}{b}$ increase from $b = 0$ to $b = \lfloor t/2 \rfloor$ and decrease from $b = \lceil t/2 \rceil$ to $b = t$. Therefore (recalling $\#W = t!$)

$$\frac{\#W}{M} = \frac{1}{t}\binom{t}{\#B} \geqslant \frac{1}{t}\binom{t}{\lceil t/3 \rceil}.$$

Finally, using Stirling's approximation we can estimate

$$\frac{1}{t}\binom{t}{\lceil t/3 \rceil} = \Theta\left(\frac{(3/2^{2/3})^t}{t^{3/2}}\right).$$

The theorem now follows from Proposition 6. $\square$

When $\Lambda = \{1\}$, $L$ is the set of all permutations over $\mathbb{A}_t$, and we recover Theorem 30 from Ellul et al.

When $\Lambda$ is either $\varnothing$ or $\{0\}$, there is a constant size context-free grammar for $L$. When $\Lambda = \mathbb{N}$, there is a context-free grammar of linear size.

Note that the language $L$ is not necessarily context-free. Parikh's theorem implies that $L$ is context-free if and only if $\Lambda$ is eventually periodic, in which case it is in fact regular.

## Acknowledgements

This note was inspired by several questions on http://math.stackexchange.com by the user `jerr18`.

We thank one of the reviewers for extremely detailed comments that greatly improved the presentation, and for bringing reference [1] to our attention.

## References

[1] K. Ellul, B. Krawetz, J. Shallit, M.-w. Wang, Regular expressions: new results and open problems, J. Autom. Lang. Comb. 10 (2005) 407–437.
[2] M. Charikar, E. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, A. Shelat, The smallest grammar problem, IEEE Trans. Inform. Theory 51 (2005) 2554–2576.
[3] J. Arpe, R. Reischuk, On the complexity of optimal grammar-based compression, in: IEEE Data Compr. Conf., pp. 173–182.
[4] P.R.J. Asveld, Generating all permutations by context-free grammars in Chomsky normal form, Theoret. Comput. Sci. 354 (2006) 118–130.
[5] P.R.J. Asveld, Generating all permutations by context-free grammars in Greibach normal form, Theoret. Comput. Sci. 409 (2008) 565–577.
[6] P.R.J. Asveld, Generating all circular shifts by context-free grammars in Chomsky normal form, J. Autom. Lang. Comb. 11 (2006) 147–159.
[7] P.R.J. Asveld, Generating all circular shifts by context-free grammars in Greibach normal form, Internat. J. Found. Comput. Sci. 18 (2007) 1139–1149.
[8] A. Kelemenová, Complexity of normal form grammars, Theoret. Comput. Sci. 28 (1983) 299–314.